DOCUMENT RESUME

ED 169 072 TM 007 827

AUTHOR Stenner, A. Jackson; And Others
TITLE The Standardized Growth Expectation: Implications for Educational Evaluation.
PUB DATE [Mar 78]
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978) ; Best copy available

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; *Achievement Rating; Achievement Tests; Compensatory Education Programs; Elementary Secondary Education; *Expectation; Grade Equivalent Scores; *Norm Referenced Tests; Research Methodology; Standardized Tests; Student Testing; Test Interpretation; *Time Factors (Learning)
IDENTIFIERS *Standardized Growth Expectation

ABSTRACT

Three assumptions underlying the use of norm referenced tests are examined: (1) that expressing treatment effects in a standard score metric permits aggregation of effects across grades; (2) commonly used standardized tests are sufficiently comparable to permit aggregation of results across tests; and (3) the summer loss of achievement observed in Title I projects is due to an actual loss in achievement and skills. Hypotheses regarding the standardized growth expectation (SGE) are also presented; SGE refers to the amount of growth (expressed in standard deviation form) that a student must demonstrate over the treatment interval to maintain standing in the norm group. SGE may also be conceptualized as the difference between the pretest percentile and the posttest percentile. Hypotheses are presented regarding the decrease in SGE which accompanies grade increases, and the variation in SGE according to the achievement test used. Further research topics investigating the validity of the SGE phenomenon are suggested. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED169072

3/23/78

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jack Stenner

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE STANDARDIZED GROWTH EXPECTATION: IMPLICATIONS FOR EDUCATIONAL EVALUATION

A. Jackson Stenner
NTS Research Corp.

Earl L. Hunter
Washington, D.C. Public Schools

June D. Bland
Washington, D.C. Public Schools

Mildred L. Cooper
Washington, D.C. Public Schools

Presented at The Annual Meeting of The
American Educational Research Association
Toronto, Canada 1978

Session 17.20
1978

TM 007 827

This paper is based on work at NTS Research Corp. conducted for the Assistant Superintendent for Research and Evaluation, Washington, D.C. Public Schools (Contract No. 0334 AA NS 07 GA).

The purpose of this paper is to review some assumptions underlying the use of norm-referenced tests in educational evaluations and to provide a prospectus for research on these assumptions as well as other questions related to norm-referenced tests. Specifically, the assumptions which will be examined are (1) expressing treatment effects in a standard score metric permits aggregation of effects across grades, (2) commonly used standardized tests are sufficiently comparable to permit aggregation of results across tests, and (3) the summer loss observed in Title I projects is due to an actual loss in achievement skills and knowledge. We wish to emphasize at the outset that our intent in this paper is to raise questions and not to present a coherent set of answers.

Throughout this paper we make use of an index termed the "standardized growth expectation" (SGE). The SGE is defined to be the amount of growth (expressed in standard deviation form) that a student must demonstrate over a given treatment interval to maintain his/her relative standing in the norm group (Stenner et al., 1977). The SGE rests on the assumption that a student will attain the same raw score on the pretest and posttest if no learning has taken place between testings. If the pretest raw score is equivalent to a national percentile of 50 and the same raw score is entered into the corresponding posttest percentile table, the resulting percentile score will be less than 50. The difference between the pretest percentile and the posttest percentile expressed in standard score form is termed the SGE. Stated another way, the SGE is the amount that a student at a particular pretest percentile is assumed to learn over a period of time or, conversely, the loss in relative standing that such a student would suffer if he/she learned nothing during the time period.

An example may help to clarify the procedures used to calculate the SGE. Table 1 presents a raw score to percentile conversion table for beginning of first grade and end of first grade on the Total Reading scale of the Comprehensive Test of Basic Skills, Form S. The average (50th percentile) beginning first grade student attains a raw score of 31 on Total Reading. Under the assumption that this illustrative average student learns nothing in the first grade, he/she would be expected to again obtain a raw score of 31 on the posttest. Whereas a raw score of 31 is equivalent to a beginning first grade percentile of 50, it represents an end of first grade

percentile of 9. If both percentiles are converted to Z scores and subtracted, the result is an SGE of 1.39 (i.e., the 50th percentile equals a Z score of zero, whereas the 9th percentile equals a Z score of -1.39). In other words, if an average student learns nothing about reading during the first grade, he/she would be expected to lose 1.39 standard deviation units in relative standing because that is the amount of standardized growth exhibited by the national norm group during the first grade.[1]

Grade-to-Grade Variation in SGEs

Some educational evaluations which employ norm-referenced achievement tests share a common assumption, namely that observed treatment effects (e.g., differences between standardized means of observed treatment group posttest scores and expected treatment group posttest scores) are comparable across grades. Stated another way, it has been assumed that a one-third standard deviation difference between experimental and control students' reading comprehension has the same meaning whether observed at the second, fifth, or seventh grade levels. It has also been assumed, with apparent logic, that if a special program consistently demonstrates larger treatment effects in the primary as opposed to intermediate grades, then compensatory efforts should be concentrated at the lower level. In fact, the twelve-year history of ESEA Title I documents a nationwide trend toward focusing increasing amounts of compensatory education efforts on primary students. Numerous evaluation studies have supported this movement through findings that larger treatment effects are possible with younger students. One question raised in this paper is whether or not there is a built-in bias in our evaluation methodology and/or instrumentation that insures finding more "exemplary" programs at the primary grade levels.

We raise this issue of cross grade comparisons because educational policy may rest upon the legitimacy of just such comparisons. For example, a review of seventy-three school desegregation studies concluded that the critical period for desegregation (to maximize black students' achievement) is prior to third grade (Crain and Mahard, 1977) Black students

---

[1] The SGE differs slightly depending upon where in the pretest distribution the raw score is selected to be entered into the posttest percentile distribution. This difference is of interest in its own right, but introducing the issue in the present paper would unnecessarily confuse the presentation.

desegregated beyond the third grade tend to show lower achievement test gains. Of the ten studies involving first and second graders, eight showed that desegregation produced higher achievement levels and two showed no effect. Only nine of twenty-one studies showed higher achievmenent among third and fourth grade students and for students in grades five through twelve, only sixteen of thirty-one studies showed any achievement gain attributable to desegregation. Eleven of the seventy-three studies reviewed by Crain and Mahard were not analyzed by grade. Taken at face value, these findings suggest that younger students benefit more from desegregation (in terms of achievement) than do older students. Crain and Mahard (1977) conclude: "The review of these studies is inconclusive or debatable on nearly every point except that desegregation in the early grades is superior to desegregation in the later grades" (p.19). It is precisely this kind of conclusion based upon cross grade comparisons of norm-referenced achievement tests that may be invalid.

Table 2 presents standardized growth expectations for five commonly used norm-referenced achievement tests. The full year SGEs show a consistent decrement with each grade. The negative relationships between grade and SGE hold for both reading and mathematics across all five tests. The differences between second grade SGEs and eighth grade SGEs averaged across tests exceeds one-half standard deviation. Interestingly, the largest losses in SGEs for both Total Reading and Total Math occur during the third grade period. Several recent studies on test score decline (cf Wirtz, 1977) have concluded that test scores begin to drop at the fourth grade level. It might be rewarding to investigate the possibility that SGE decrements are causally implicated in reported test score declines. The Stanford Achievement Test, for example, exhibits almost a fifty percent decrement in Total Reading SGE from second to third grade. Similarly, the ITBS, CAT-77, and CTBS-S all show decrements approaching twenty-five percent.

Insight into the implications that these grade-to-grade differences may have for educational evaluation is gained by realizing that a treatment effect of one-third standard deviation (often employed as a threshhold value for an educationally meaningful or practically significant effect size) represents a 33% increase above expectation for second graders on the MAT Total Reading and a 200% increase above expectation for eighth graders. If the ongoing instructional process is incapable of producing more than .15 SD's of growth on MAT Total Reading among eighth graders, then it seems somewhat unrealistic to expect an eighth grade compensatory program or desegregation effort to demonstrate a treatment effect of .33 SD's.

The conclusion implied above is that any cross grade comparisons of treatment effects expressed as national percentiles, standard scores, NCEs or grade equivelents are usually inappropriate. Although treatment effects expressed in standard score form may be smaller at the eighth grade level than at the second grade level, the statistical significance (e.g., F ratio) of these effects may be the same for both grade levels (assuming equal sample sizes) because of the increased pre-post correlation at the eighth grade level. Thus to apply an arbitrary treatment effect criterion of .33 SD's (or any other uniformly applied criterion based upon standardized scores) when screening for exemplary projects or reviewing research studies, unfairly discriminates against upper grade projects. A metric for practical or educational significance which would be comparable across grades cannot be formulated without consideration of the fact that pre-post correlations increase with grade.[1]

Following are five hypotheses regarding decrement in SGE as grade increases. It is higly likely that several of these alternative explanations combine to account for cross grade differences. Although the first two hypotheses are intuitively more appealing than the others, much more study of the merits of each explanation is recommended.

Domain Expansion Hypothesis: *With each increase in grade the relevent domain (e.g. reading or math) expands in terms of the number of concepts encompassed by the domain.* The result of an expanding domain is that a fixed number of items will be less and less representative; proportionately fewer items can be allocated to any given span of concepts and objectives. As the range of concepts and objectives covered by a test increases, the SGE decreases and edumetric validity is reduced (cf Carver, 1974). The poorer the match between what is taught at a given grade level and what is tested, the less sensitive the test is to growth, and the lower the SGE.

Shifting Constructs Hypothesis: *The levels of some tests are not well articulated and with each succeeding grade stable organizing influences other than reading or math achievement increasinly determine students' scores on norm-referenced tests.* For example, if reasoning ability becomes

[1] Perhaps a more methodologically defensible metric would be the standard deviation of the pre-post residuals. This metric should be comaprable across grades since it is adjusted for pre-post correlations.

progressively more confounded with reading and math achievement scores as greade increases, and reasoning ability grows at an increasingly lower rate than reading and math achievement, then the confounded reading and math SGEs would be expected to decline as confounding increases. As what is measured by norm-referenced reading and math tests changes, the edumetric validity of these tests may be reduced.

Learning Curve Hypothesis: *The deteriorating SGE is due to an actual slowing in the rate of learning similar to the way height slows down from birth to eighteen years of age.* According to this hypothesis younger students have a greater capacity for learning and this capacity deteriorates with age.

Unequal Interval Hypothesis: *Standard deviation units are not equal interval across grade.* Imagine a rubber band marked into ten equal intervals representing the one standard deviation SGE at second grade on the MAT Total Reading. Now imagine the rubber band stretched to the point that the distance between any two marks is equal to the entire length of the unstretched rubber band. In this way we can see how growth at the seventh grade on the MAT Total Reading (SGE = .10) might equal growth at the second grade (SGE = 1.0). If this hypothesis were accepted, the validity of cross grade comparisons would be questionable, but so would just about all other comparisons of interest in educational evaluation.

Instructional Emphasis Hypothesis: *Upper grade teachers do not emphasize reading and math as much as lower grade teachers and, as a consequence, students learn less and subsequently show less growth on norm-referenced reading and mathematics tests.* As upper grade teachers concentrate less on reading and math instruction than primary grade teachers, the SGE decreases.

The five hypotheses are rank ordered from most likely to least likely (in our opinion) as explanations for the observed decrement in SGE as grade increases. The first two hypotheses state that as grade increases, the edumetric validity of NRTs decreases. The second three hypotheses offer explanations which, although not related to the edumetric properties of NRTs, cannot be discounted without further research. At present all the hypotheses and the rank ordering are exercises in speculation. However, we are confident

that variation in the SGE across grades represents an important phenomenon which may have implications for both policy makers and evaluation specialists. Until the grade-to-grade fluctuations in SGE are better understood, researchers might refrain from sweeping policy recommendations based upon cross grade comparisons of norm-referenced achievement test scores.

Test-to-Test Variation

Almost as striking as the grade-to-grade variations in full year SGEs within a test are the test-to-test differences within a grade. Examination of Table 2 reveals numerous instances of SGEs being thirty to forty percent higher for some tests than for others. When we shift focus to school-year SGEs (see Table 3), the differences across tests are even more dramatic. School-year SGEs frequently vary among tests by as much as fifty to sixty percent with isolated instances of SGEs for some tests being three to six times as large as those of other tests.

All other things being equal, the higher the match between what is learned and what is tested (i.e., the higher the edumetric validity) the higher the SGE. An SGE near zero means that either nothing was taught, or something was taught but nothing was learned, or the test did not reflect what was taught and/or learned. A large SGE suggests that something was learned and the test reflects well whatever was learned. Presumably, criterion-referenced tests are superior to norm-referenced tests precisely because they provide a better match between what is taught/learned and what is tested. The SGE may provide a simple index for evaluating the claims made on behalf of criterion-referenced tests that they are superior evaluation tools. If CRTs demonstrate higher SGEs than NRTs, then these claims are likely valid. (The last section on the edumetric ratio addresses this issue more thoroughly). A properly developed CRT should have greater fidelity to the curriculum and, consequently, larger SGEs. The SGE may be an effective means of assessing, a priori, tests' probably sensitivity to instruction.

We offer four hypotheses for the variation in SGEs across tests. Again we order the hypotheses in terms of our present thinking regarding the probability that each hypothesis will be sustained in future studies.

Edumetric Hypothesis: *Norm-referenced tests differ in the extent to which they reflect stable between-individual differences (Carver's psychometric dimension) and the extent to which they reflect within-individual growth (Carver's edumetric dimension).* A test may possess exemplary psychometric properties (e.g., hign internal consistency and a good p value distribution) but be insensitive to what students learn over a given treatment interval. Such a test will have a low SGE but be otherwise indistinguishable from other norm-referenced tests. The reader is encouraged to re-examine Table 2 in light of this hypothesis.

Procedures Hypothesis: *Test publishers use vastly different approaches to interpolation/extrapolation and make different assumptions regarding summer growth, thus artificially creating SGE differences.* The fact that full year SGEs are much more comparable between tests than are school year or summer SGEs suggests that publishers differ considerably in the assumptions they make about summer growth.

Norm Group Hypothesis: *The composition of the norm groups for the various tests differ to such an extent that the SGEs are affected.* Suppose, for example, that the Stanford Achievement Test (SAT) norm group was substantially brighter than the Metropolitan Achievement Test (MAT) norm group. The result would be that the Stanford Achievement Test norms would reflect more growth and, consequently, the SGEs for the SAT would be larger than those for the MAT. We should note that findings from the Anchor Test Study, for at least four of the tests considered in this paper, do not account for the large SGE differences across tests.

Cohort Hypothesis: *Although the norm groups for the various tests were selected in essentially similar ways because the tests were normed in different years, the samples may have differed in rate of achievement.* Teachers are fond of claiming that, like fine wines, there are "vintage years" in which a particular group of students just seems brighter, however, the pattern of SGE differences across tests (taking into consideration the year each test was normed) is not consistent with this hypothesis.

Of the four hypotheses just presented, the procedures and edumetric hypotheses seem most compelling. The fact that full-year SGEs are substantially more comparable across tests than either school-year or summer SGEs suggests

that publishers may make different assumptions about what students learn during the summer period. Apparently publishers of the Stanford Achievement Test assume that very little reading or math achievement growth should be expected of a fiftieth percentile student, whereas publishers of the MAT seem to assume a large amount of summer growth.[1] It seems probable that evaluation findings will vary depending upon which test is used, how closely different publishers' assumptions regarding summer growth coincide with empirical findings, and whether fall to spring or spring to spring testing dates are employed.

According to the edumetric hypothesis, some norm-referenced tests are more sensitive to student growth in reading and math than other tests. Those tests with low SGEs measure well the between-individual differences which become more and more stable as students get older, but do a relatively poor job of measuring what students learn during a particular treatment interval. Most users of NRTs, particularly evaluation specialists, are primarily interested in measuring achievement growth. The SGE differences across tests seem to indicate that commonly available NRTs differ considerably in their edumetric validity, i.e., sensitivity to instruction.

The implications for educational evaluation of sustaining the edumetric hypothesis are substantial indeed. First of all, assuming the validity of this hypothesis, it is little wonder that most of our school effects studies have accounted for such minuscule proportions of variance with instructional process measures (cf Cooley and Lohnes, 1976). The problem may not rest with so-called "weak treatments" but rather with measurement instruments that are systematically biased against showing either significant treatment-control differences or substantial process-outcome relationships. When the SGE is as small as .15 standard deviations, as is the case with several tests at the eighth grade level, is it any wonder we find very few "exemplary" eighth grade reading and math programs or that Coleman (et al. 1966) could find so few school variables that correlated with STEP Reading Test scores. Similarly, it is perhaps no coincidence that at those grade levels where the SGEs are largest and, presumably the edumetric validity of the tests is highest, we find a higher frequency of "exemplary" projects. An evaluation

[1] The fact that both the Stanford and Metropolitan claim to have empirically determined fall and spring norms makes the substantial differences in summer SGEs for these two tests all the more puzzling.

study that employs an NRT with a low SGE may be a priori doomed to add yet another conclusion of "no significant difference" to the literature on school effects.

Summer Loss Phenomenon

Several recent studies have highlighted the fact that Title I students achieve above expectation during the regular school year and lose in relative standing during the summer months (Pelavin and David, 1977; Stenner et al., 1977). Title I projects that use fall to spring testing dates often report substantial treatment effects, whereas projects that use fall to fall or spring to spring testing dates often report no treatment effects (Pelavin and David, 1977). In general, there has been limited appreciation for the different conclusions regarding treatment effects that result from simply varying testing dates. For example, tentative procedures in the OE Title I Evaluation System call for aggregating treatment effects without regard for testing dates. Similarly, the Joint Dissemination Review Panel typically evaluates reported treatment effects without considering testing dates.

Table 4 presents standardized growth expectations for the summer period (spring to fall). Except for the SAT, all tests exhibit substantial growth expectations over the summer period. We suggest that an edumetrically valid achievement test should have a large SGE over the school year and a small summer SGE. However, since the size of both school-year and summer SGEs can be manipulated by making different assumptions about summer growth, the data presented in this paper cannot speak directly to this point. If empirical data could be collected at three points in time (fall, spring, fall) for all commonly used NRTs, then the ratio of summer growth to school-year growth might address the question of comparative edumetric validity. Under such an analysis, when SGEs for the summer period approach or exceed SGEs for the school year, a test's sensitivity to instruction must be questioned. Large summer SGEs would suggest that the construct being measured by the test evidences growth whether or not the student is in school. Such an instrument would not only be relatively less sensitive to instruction-related achievement growth but would also presumably be insensitive to special project treatment effects. Again we emphasize that given the lack of multiple

empirical norming points, the summer - school-year ratios given in Table 5 may simply reflect variation in publishers' assumptions about summer growth.

The large summer SGEs exhibited by four of the five tests examined in this paper raises questions about how much of the report summer loss among Title I student is due to absolute loss in achievement and how much is due to assumptions made by publishers. If Title I students actually lose raw score points over the summer period then we must conclude that there is an absolute loss in acquired skills and knowledge. If, however, there is no raw score change from spring to fall, then the Title I summer loss is relative rather than absolute and is a function of publisher assumptions. Discussions with other researchers studying this phenomenon suggest that there is some doubt as to whether the absolute achievement loss among Title I students is as large as is commonly believed. According to the arguments presented in this paper, the amounts of both absolute and relative loss may depend upon the NRT employed in the evaluation.

The Edumetric Ratio

The fact that students do not attend school year around suggests a means for computing edumetric validities for commonly used norm-referenced and criterion-referenced tests. An edumetrically valid test, i.e., a test which is sensitive to instruction, should evidence proportionately higher SGEs during the school year than during the summer. If we assume a nine month school year and a three month summer, then any test purporting to measure what is taught in school (e.g., reading comprehension and math computation) should evidence a ratio of "school-year SGE" to "summer period SGE" larger than 3:1 (For convenience, we term this value the edumetric ratio). On the other hand, a test of nonverbal reasoning might evidence an edumetric ratio near 3:1, indicating that noverbal reasoning (or what Cattell (1971) calls Fluid Ability) grows at a constant rate largely unaffected by school experiences.[1] Tests purporting to measure skills and objectives taught in school which show edumetric ratios near "three" would probably prove highly insensitive to treatment effects and might be expected to evidence near zero correlations with variables similar to those employed by Coleman et al. (1966).

[1] Edumetric ratios computed separately for different socio-economic groups might provide insight into how differences in out-of-school and in-school experiences impact on achievement.

The edumetric ratio may also provide a means of externally validating criterion-referenced test items. Typically CRT validation efforts rely heavily on content analysis and judgements of curriculum experts regarding the match between curriculum and what a test item presumably measures. Edumetric validity of such items is assumed when judges agree on what concept or objective an item is measuring. We suggest that rating consensus is insufficient evidence to conclude that a test item is edumetrically valid. One more methodologically defensible approach might be to compute edumetric ratios on a set of items judged to be measuring a particular concept or objective and include on the final instrument only those items with high ratios.

Lastly, a comparison of SGEs for a widely used achievement and ability test offer some additional insights into the aptitude-achievement distinction (Green, 1974). Judging from theory and publisher test descriptions, one would expect achievement tests to have higher SGEs than ability tests. For example, the Technical Manual for the Cognitive Abilities Test (Thorndike and Hagen, 1971) states: "...The test can be characterized by the following statements and these characteristics describe behavior that is important to measure for understanding an individual's educational and work potential: (1) The tasks deal with abstract and general concepts, (2) In most cases, the tasks require the interpretation and use of symbols, (3) In large part, it is relationships among concepts and symbols with which the examinee must deal, (4) The tasks require the examinee to be flexible in his basis for organizing concepts and symbols, (5) Experience must be used in new patterns, and (6) Power in working with abstract materials is emphasized, rather than speed" (p.25). Contrast the above description with that given in the technical manual for the Iowa Test of Basic Skills, ..."The ITBS provides for comprehensive and continous measurement of growth in the fundamental skills: vocabulary, reading, the mechanics of writing, methods of study, and mathematics. These skills are crucial to current day-to-day learning activities as well as to future educational development" (p.3). In the ability test manual, phrases such as "educational potential," "general concepts," and "interpretation and use of symbols" are used whereas the achievement test manual uses such terms as "growth," "fundamental skills," "diagnosis," and "skill improvement." Clearly the impression one gets from these two manuals is that the Cognitive Abilities Test measures something more stable and less sensitive to school experiences than the ITBS; an impression which is not sustained by the SGE data.

Table 7 contrasts SGEs for the Cognitive Abilities Test and the ITBS. A first observation is that ITBS-Reading SGEs are comparable to CAT-Verbal SGEs. Thus, the ITBS-Reading appears to be almost as sensitive to instruction as the CAT-Verbal. Whether comparability between the two is due to the fact that the achievement test is actually more an ability test or the ability test is just a relabeled achievement test, or the distinction between verbal ability and reading is a sham, merits further study.[1] One conclusion appears disconcertingly clear, the ITBS-Reading appears to be only slightly more edumetrically valid than the CAT-Verbal. How serious this predicament is depends on whether one elects to fault the CAT for being too much like an achievement test or the ITBS for being too much like an ability test.

The CAT-Quantitative appears to be less edumetrically valid than the ITBS-Total Math, but more sensitive to instruction than the CAT-Nonverbal. Since the CAT-Quantitative items loaded highly on the nonverbal factor and failed to define a quantitative factor (Thorndike and Hagen, 1971, p.32) one is left with the possibility that the Quantitative items are simply a mixture of items similar to ITBS-Total Math items and nonverbal reasoning items. Had the Quantitative Scale held more true to its label, we suspect that the SGEs would more closely approximate those for ITBS-Total Math. Finally, the CAT-Nonverbal evidences the lowest SGE. Whether the nonverbal growth expectations for the summer period are proportional to the school year, indicating little school effect, is a question requiring further study.

A Prospectus For Research

A major thesis of this paper is that policy decisions based upon grade-to-grade and test-to-test comparisons rest on a potentially shaky foundation. If the SGE index is meaningful and the analyses based upon it are valid, then a potentially large number of research findings merit re-examination. Granting the far-reaching policy implications inherent in our assertions and the need to establish quickly whether or not these assertions are valid, we offer the following research agenda.

---

[1] We are not suggesting that just because two tests have similar SGEs they are necessarily measuring the same thing. We are suggesting that evidence of comparable SGEs when added to information that disattenuated inter-test correlations approach 1.00 provides a pretty strong case for the fact that the two tests measure the same psychological construct.

- Submit the SGE concept to comprehensive analysis by measurement specialists focusing upon the conceptual basis for the index and assumptions underlying its computation.

- Compute SGEs for all subtests of commonly used achievement and ability tests marketing during the past twenty years, and compare SGEs across grades and subtests. Some form of multi-method, multi-trait analysis might prove useful in such a substudy (suggested by Tony Conger: personal communication).

- Conduct a comprehensive content analysis of commonly used NRTs to determine the extent to which item content, type, or format contribute to variability in SGEs across tests (suggested by Joe Haenn: personal communication).

- Conduct a meta analysis of reported treatment effects across a wide range of studies to determine whether treatment effects are correlated with SGEs. Preliminary investigations suggest that this may be a particularly fruitful area for further investigation.

- Conduct a logical and empirical analysis of the summer loss phenomenon found among Title I students. Estimate, if possible, what proportions of the loss are relative and absolute, and examine ways these proportions differ depending upon which NRT is used. Also conduct an item analysis to determine which skills evidence the largest losses over the summer.

- Conduct a preliminary investigation of the relationship between shifting edumetric validity and the Scholastic Aptitude Test score decline.

- Compute SGEs for a sample of criterion-referenced tests and investigate the claim that the SGE provides a useful index for comparing edumetric validities of CRTs and NRTs.

- Conduct a logical and empirical examination of the effects of out-of-level testing on the edumetric validity of NRTs.

The above research agenda will first address the utility and validity of the SGE concept and then proceed to examine selected implications of sustaining the edumetric hypothesis. The current nationwide interest in basic skills testing makes the topic of the proposed research particularly policy relevant at this time.

References

Carver, Ronald P. Two Dimensions of Tests - Psychometric and Edumetric. American Psychologist, 1974, 512-518.

Cattell, R. B. Abilities: Their Structure, Growth and Action. Boston: Houghton-Mifflin, 1971.

Coleman, J. S. et al. Equality of Educational Opportunity. Washington, D. C., U.S. Office of Education, 1966.

Cooley, William W. and Lohnes, Paul R. Evaluation Research in Education. New York: Irvington Publishers, Inc. , 1976.

Crain, Robert L. and Mahard, Rita E. "Desegregation and Black Achievement." National Review Panel on School Desegregation: Amelia Island, Florida, 1977.

Green, Donald R. The Aptitude-Achievement Distinction. Monterey, California, CTB/McGraw Hill, 1974.

Pelavin, Sol H. and David, Jane. "An Analysis of Longitudinal Data from Compensatory Education Programs", Stanford Research Institute: Menlo Park, California, 1977.

Stenner, A. Jackson; Strang, Ernest W.; and Baker, Robert F. "Technical Assistance in Evaluating Career Education Projects - Volume II." DHEW/Office of Education, Contract No. 300760312, 1978 (in press).

Stenner, A. Jackson; Riegel, N. Blyth; Feifs, Helmuts A.; and Davis, B. Steven. Evaluation of The ESEA Title I Program of The Public Schools of The District of Columbia: Washington, D.C., 1977.

Thorndike, R. L. and Hagen, E. Cognitive Abilities Test. Boston: Houghton Mifflin, 1971.

TABLE 1

RAW SCORE TO PERCENTILE TABLE FOR BEGINNING AND END OF FIRST GRADE ON CTBS, LEVEL B

Total Reading

Beginning of First Grade		End of First Grade	
Raw Score	Percentile	Raw Score	Percentile
73-84	99	84	99
86-72	98	84	98
65-67	97	84	97
61-64	96	84	96
59-60	95	84	95
57-58	94	83	94
55-56	93	83	93
53-54	92	82	92
52	91	82	91
31	50	59	50
31	49	58	49
31	48	58	48
31	47	57	47
31	46	56	46
30	45	55	45
29	44	54	44
29	43	53	43
29	42	53	42
29	41	52	41
20	10	32	10
19	9	31	9
18	8	30	8
18	7	29	7
18	6	28	6
18	5	27	5
17	4	25-26	4
16	3	24	3
15	2	21-23	2
0-14	1	0-23	1

TABLE 2
STANDARDIZED GROWTH EXPECTATIONS FOR SELECTED TESTS
AND GRADES (50%TILE) .
(STANDARD DEVIATION UNITS)

	TOTAL READING					TOTAL MATH				
Spring to Spring Grade Period	Stanford [1]	ITBS[2]	CAT 77 [3]	CTBS-S [4]	Metropolitan [5]	Stanford	ITBS	CAT-77	CTBS	Metropolitan
1.7 - 2.7	1.22	.95	1.08	1.04	?	1.17	1.17	1.23	1.17	?
2.7 - 3.7	.64	.74	.74	.74	.71	.99	,84	.99	.99	.84
3.7 - 4.7	.56	.74	.61	.52	.71	.58	.91	.71	.56	.88
4.7 - 5.7	.44	.74	.47	.41	?	.64	.71	.64	.61	?
5.7 - 6.7	.33	.57	.36	.30	.44	.44	.67	.49	.36	.58
6.7 - 7.7	.30	.47	.30	.30	.30	.33	.49	.36	.30	.47
7.7 - 8.7	.30	.41	.38	.28	.41	.41	.47	.47	.30	.38
8.7 - 9.7	.20	.33	.25	.23	?	.28	.33	.28	.33	?

[1] Stanford was normed in October and May
[2] ITBS was normed in November
[3] CAT-77 was normed in November and May
[4] CTBS was normed in April
[5] Metropolitan was normed in October and April

? - Standard procedure for computing SGE could not be followed given that spring to spring norms are not available for indicated levels of the MAT

TABLE 3
STANDARDIZED GROWTH EXPECTATIONS FOR SELECTED TESTS
AND GRADES (50%TILE)
(STANDARD DEVIATION UNITS)

	TOTAL READING					SCHOOL YEAR	TOTAL MATH				
Fall to Spring Grade Period	Stanford	ITBS	CAT-77	CTBS	Metropolitan		Stanford	ITBS	CAT-77	CTBS	Metropolitan
2.1 - 2.7	.95	.74	.56	.51	1.00		.92	.92	.77	.61	1.17
3.1 - 3.7	.58	.47	.49	.41	.36		.84	.58	.84	.55	.84
4.1 - 4.7	.49	.41	.30	.30	.30		.58	.44	.52	.30	.58
5.1 - 5.7	.38	.38	.25	.28	.30		.64	.44	.44	.33	.30
6.1 - 6.7	.30	.33	.20	.23	.25		.33	.38	.33	.31	.[illegible]
7.1 - 7.7	.28	.28	.18	.18	.10		.30	.28	.25	.15	.05
8.1 - 8.7	.18	.28	.18	.18	.15		.25	.30	.18	.15	.10
9.1 - 9.7	.18	.12	.10	?	?		.20	.20	.12	?	?

? - Could not compute these values from information given in publishers' manuals.

TABLE 4
STANDARDIZED GROWTH EXPECTATIONS (50%TILE) FOR SELECTED TESTS AND GRADES
(STANDARD DEVIATIONS)

Spring to Fall Grade Period	TOTAL READING					TOTAL MATH				
	Stanford	ITBS	CAT-77	CTBS	Metropolitan	Stanford	ITBS	CAT-77	CTBS	Metropolitan
2.7 - 3.1	.10	.38	.23	.36	.41	.15	.15	.23	.52	.36
3.7 - 4.1	.02	.25	.30	.23	.36	.05	.38	.25	.23	.25
4.7 - 5.1	.02	.23	.20	.15	?	.00	.28	.25	.28	?
4.7 - 5.1	.05	.25	.15	.12	.18	.10	.25	.18	.20	.30
6.7 - 7.1	.00	.20	.15	.12	.25	.02	.23	.12	.20	.25
7.7 - 8.1	.10	.15	.20	.10	.36	.15	.20	.28	.15	.36
8.7 - 9.1	.02	.10	.15	?	.15	.20	.15	.20	?	.10

? - Could not compute these values from information in the CTBS manuals.

TABLE 5
RATIO OF SUMMER GROWTH EXPECTATION TO SCHOOL YEAR EXPECTATION (50%TILE)*

	TOTAL READING					TOTAL MATH				
Grade	Stanford	ITBS	CAT-77	CTBS	Metropolitan	Stanford	ITBS	CAT-77	CTBS	Metropolitan
2	.11	.51	.41	.70	.41	.16	.16	.29	.85	.30
3	.03	.53	.62	.56	1.00	.06	.66	.30	.42	.30
4	.04	.56	.65	.50	?	**	.64	.48	.93	?
5	.13	.56	.60	.42	.60	.16	.57	.40	.67	1.00
6	**	.61	.76	.52	1.00	.06	.61	.37	.65	1.67
7	.36	.54	1.13	.56	3.60	.50	.71	1.11	1.00	7.20
8	.11	.36	.86	?	1.00	.18	.50	1.14	?	1.00

*As an illustration the growth expectation over the summer interval on the CTBS Reading scale (for second graders) is 70% of the regular school year growth expectation for a 50th percentile student.

**Growth expectation over the summer period is zero.

? - Could not compute this value from information provided in the MAT Manual.

TABLE 6

STANDARDIZED GROWTH EXPECTATIONS FOR THE IOWA TEST OF BASIC SKILLS AND THE COGNITIVE ABILITIES TEST

	ITBS Reading Comprehension	Cognitive Abilities Test Verbal	ITBS Total Math	Cognitive Abilities Test Quantitative	Cognitive Abilities Test Nonverbal
3.7 - 4.7	.74	.74	.91	.71	.38
4.7 - 5.7	.74	.61	.71	.52	.38
5.7 - 6.7	.57	.52	.67	.38	.23
6.7 - 7.7	.47	.41	.49	.36	.20
7.7 - 8.7	.41	.30	.47	.36	.23
8.7 - 9.7	.33	.30	.33	.28	.23